EPID708: Machine Learning for Epidemiologic Analysis in the Era of Big Data

July 7-11, 2025 / 1:30-5:00 p.m. EDT

Remotely via Zoom

Professor: David McCoy
david_mccoy@berkeley.edu

Course Description

Course focuses on advances in machine learning and its application to causal inference and prediction via Targeted Learning, which allows the use of machine learning algorithms for prediction and estimating so-called causal parameters, such as average treatment effects, optimal treatment regimes, etc. We will discuss implementation via cloud computing.

Course Materials

M. van der Laan and S. Rose. Targeted learning: causal inference for observational and experimental data. Springer, 2011.

Pre-requisites

Introductory course in statistics as well as courses or working knowledge of basic regressions (linear, logistic, etc.). Having some background in the programming language R preferred.

Course Goals

- A basic understanding of causal inference, including structural causal models, definition of causal parameters via counterfactual distributions, and ways to establish identifiability from observed data.
- Familiarity and ability to implement machine learning, specifically the concepts of SuperLearning and the power of cross-validation in data-adaptive estimation.
- Ability to apply machine learning algorithms to prediction problems and estimate and derive inference for the resulting fit.
- Ability to use the fits of machine learning algorithms to estimate causal effects using simple substitution estimators.
- Ability to apply Targeted Learning approaches (e.g., targeted maximum likelihood estimation) to estimate, using machine learning, a priori specified treatment effects as well as general variable importance measures.

- A basic understanding of how to use parallel computing and large computer clusters to be able to estimate using computer intensive algorithms on large (Big Data) data sets.
- How the general methodology applies to goals of Precision Medicine.

Competencies

- Ability to apply estimation roadmap to novel data questions.
- Ability to implement estimation via R and existing software packages.
- Basic knowledge of how to use such algorithms on Big Data including the use of cloud computing.

We will be doing high level statistics and using machine learning to estimate more sophisticated measures of association, but related to the standard measures taught in statistics for epidemiology. (Epidemiology MPH Core Competency #3)

We will be discussing estimation of causal effects in the context of observational and randomized trials. We also will teach general approach for deriving estimates from any potential biased and unbiased sample (via the estimation roadmap). (Epidemiology MPH Core Competency #5)

We will discuss identification of causal effect from either nonparametric structural equation models as well as DAG's. Students will have some experience in deriving statistical estimates for causal quantities under assumptions. (Epidemiology MPH Core Competency #6)

Time will be spent on how estimation in an appropriate statistical model (that is generally with little constraints on the data-generating model) can greatly reduce bias. We will define confounding as part of our discussion of causal inference (as in #6 above). (Epidemiology MPH Core Competency #7)

Course Requirements

Class Participation	20%
Final Project	80%
Total	100%

Classroom Expectations/Etiquette

Discuss any specific issues/expectations regarding class format, classroom expectations, and etiquette/decorum. This could include expectations regarding class attendance, use of laptops in class, group work, etc.

Academic Integrity

The faculty and staff of the School of Public Health believe that the conduct of a student registered or taking courses in the School should be consistent with that of a professional person. Courtesy,

honesty, and respect should be shown by students toward faculty members, guest lecturers, administrative support staff, community partners, and fellow students. Similarly, students should expect faculty to treat them fairly, showing respect for their ideas and opinions and striving to help them achieve maximum benefits from their experience in the School.

Student academic misconduct refers to behavior that may include plagiarism, cheating, fabrication, falsification of records or official documents, intentional misuse of equipment or materials (including library materials), and aiding and abetting the perpetration of such acts. Please visit http://www.sph.umich.edu/academics/policies/conduct.html for the full SPH Code of Academic Integrity and further definition of these terms.

Student Well-being

SPH faculty and staff believe it is important to support the physical and emotional well-being of our students. If you have a physical or mental health issue that is affecting your performance or participation in any course, and/or if you need help connecting with University services, please contact the instructor or the Office of Academic Affairs.

Please visit http://www.sph.umich.edu/students/current/#wellness for more information.

Student Accommodations

Students should speak with their instructors before or during the first week of classes regarding any special needs. Students can also visit the Office of Academic Affairs for assistance in coordinating communications around accommodations.

Students seeking academic accommodations should register with Services for Students with Disabilities (SSD). SSD arranges reasonable and appropriate academic accommodations for students with disabilities.

Course Topics/Reading List

Intro and Causal Inference (Day 1)

- Motivation for Course: Machines for targeted inferences
- History of parametric models
- Introduction of machine learning - allows removing the art from data analysis
- The incompatibility of using standard techniques with estimating rigorously without arbitrary assumptions
- Going for statistical machines - given the data, causal model, parameter of interest, will automatically derive optimal estimate and robust inference (the future is now).
- Roadmap for Targeted Machine Learning

  - In order to make machines, need a rigorous roadmap so that real knowledge is used optimally, and arbitrary assumptions are kept to minimum.
  - Introduce a roadmap of estimation/inference that can be applied very generally.
- causal model/ structural equation model
- intervention on SEM - counterfactuals
- identifiability
- examples of parameters

SuperLearning/Machine Learning (Day 2-3)

- Loss/Risk
- consistent estimation of risk - Cross-validation
- Oracle Inequality
- Ensemble Learning
- SuperLearning
- Implementations for Big Data

Substitution Estimators (Day 4-5)

- General concept of substitution estimator (SubEst)
- Correspondence of coefficient and SubEst in special case of linear model
- Saturated model
- Use of machine learning for data-adaptively estimating model for SubEst's.
- Example of average treatment effect
- Variable importance ideas
- Implementation for cloud computing